Job offer – Post-doctorate in Computational Biology and Statistical Biological Physics
(*For non-French scientists only*)

Research Project Short Title as Submitted to CEFIPRA: "Deciphering RNA Evolutionary Landscape by Integrating Directed Evolution and Machine Learning (RE-LEARN)"
Principal Investigator contact (Name and email id): "Martin Weigt, Sorbonne Université - Faculté des Sciences martin.weigt@sorbonne-universite.fr"
Reference Number of the Job Offer: IFI_CEF_26_02

Project description

- **Keywords :** RNA evolutionary landscape, machine learning, biological physics, RNA language model, RNA modelling, directed evolution
- **Context** : Molecular evolution is one of the most fundamental processes in biology, being at the basis of the diversity and complexity of life as we observe it today. Inspiration drawn from natural evolution has played an important role in designing therapeutic targets, enhancing enzymes for commercial use, and developing probes for diagnosis. In this regard, generating fitness landscapes and evolutionary trajectories using directed evolution experiments provides some understanding of the evolutionary process. However, due to constraints of experimentally accessing a combinatorially huge sequence space, these protocols only explore local minima around the wild-type sequence. Recent advances in machine learning and artificial intelligence have shown potential in aiding the molecular evolution process to isolate improved variants using directed evolution (DE) protocols. Even though ML/AI algorithms have been used for understanding fitness landscapes, such an approach has never been applied to in vitro selected functional RNA and isolating novel functional RNAs.
- **Abstract of the Research Project :** Though exploiting biomolecular evolution can help design industrial enzymes and drug targets, a comprehensive understanding of the evolutionary landscape necessitates exhaustive experimental exploration of vast sequence space. This makes the entire process resource-intensive and laborious creating a bottleneck. The proposed project aims to combine generative probabilistic models and direct coupling analysis with experimental-directed evolution protocols to identify novel functional RNAs. While generative models may overlook subtle aspects of the evolutionary fitness landscape discoverable only experimentally, they have the potential to produce functional sequences that diverge from any natural sequence. Our novel approach will combine our theoretical and experimental expertise to investigate unexplored regions of the fitness landscape to identify distant but fit mutants, essentially creating non-natural RNA families from scratch. To achieve this, we will train generative models on the output of directed evolution experiments; models will in turn guide the new directed evolution experiments to explore even larger sequence space. The proposed project will first experimentally measure sequence-function relationships of 10^3 sequences to create a training dataset for the model. Using this data, model-guided directed evolution will be conducted next under various selection pressures. Sequence data from both 'fit '(functional) and 'unfit '(non-functional) sequences from each directed evolution round will be fed into generative models to create variants for the next selection cycle. This will enable us to find new functional sequences and investigate the 'inaccessible 'areas of the fitness landscape. Furthermore, the iterative nature of modelling and experimentation and the feedback from experimental functional annotations will facilitate the development of an integrative generative model.
- **Scientific Objectives of the Project :** We aim to explore the large sequence space to identify novel functional RNA sequences by carrying out ML/AI-guided directed evolution experiments for functional RNA under selection. We will iterate experiments and theoretical modelling where models are trained on DE data which then guide new DE experiments exploring larger and larger sequence spaces, providing better data for models etc. We will use a green-fluorescent RNA as experimental system. Though a product of a DE protocol, this RNA has a conserved RNA G-quadruplex structure (rG4s) at its core. The rG4s is a prominent feature of the natural sequence space: its dysfunction has been implicated in several pathological conditions, and rG4s have been found to be enriched in several viral infections. Therefore, learning the detailed evolutionary landscape will not only enrich our understanding of the evolution of functional RNAs but will also help in generating novel datasets for rG4s which can be used to develop rigorous algorithms to predict and search rG4s in the natural sequence space. Furthermore, the proposed project will

also address one of the pertinent issues in functional sequence design: finding novel sequences which are very distant from WT (>40 mutations) but still functional.

- **Methodology and Timeline of the Project :** The planned research project will be carried out systematically over the course of 36 months, with four distinct tasks. Task 1. Creating initial training datasets by experimentally mapping sequence-function relationship of ~1000 RNA variants and building generative model development based on them. Task 2. Once the first generative models are developed, machine learning-guided DE experiments will be carried out starting from a much bigger RNA sequence library (~1e18). Task 3. The identified sequences from task 2 will be further characterised for their fitness and RNA structure. Task 4. Learning from the evolutionary landscape of functional RNA from DE experiments, the generative models will be further improved and the comprehensive fitness landscape will be modelled. The postdoctoral candidate to be hired for Sorbonne University will take care of the computational modelling aspects of the project, in close collaboration with the experiments performed at the Indian partner lab at Ashoka University.

Candidate profile

- Candidates can be all nationalities except French. In case of double nationality (French and another one), the candidate is not eligible. In the context of CEFIPRA, Indian candidates are preferred
- Applicants for post-doctorate must have a PhD degree (or be in the process of obtaining one) ;
- No competences in French language is required
- Candidate competences: We are looking for a highly motivated candidate with a strong theoretical or computational background in at least one of the following areas: statistical and biological physics, computational biology, machine learning applied to biological data. A PhD in a theoretical or computational subject is needed. The candidate is expected to show high autonomy and motivation for communicating with the experimental partners, within a highly interdisciplinary setting.
- Candidate know-how: The candidate is expected to have a proven knowledge in the computational modelling in statistical and biological physics (e.g. biopolymers, soft matter, disordered systems, complex systems) and/or in advanced statistical and machine learning techniques. Prior experience in modelling RNA and other nucleic acids is not necessary but a plus in the application. The candidate is not required to perform experiments.
- Expected starting date: April - May 2026
- Expected duration: 20 months (12 months of initial scholarship + 8 months of potential extension)

How to candidate ?

Documents to be provided :

i. A cover letter (reasons for the candidature, professional project …) max 2 pages
ii. A copy of the master's degree or a proof of the program followed (and expected date of end) OR A copy of the PhD degree or a proof of the PhD program followed (and expected date of defense) max 1 page
iii. A copy of results for previous scholarship (max 3 pages)
iv. A copy of Passport
v. International curriculum vitae (max 2 pages)
vi. Two letters of recommendation: one from any Indian institution and one from the French institution planned to host the candidate –mandatory- (max 2 pages)
vii. All should be submitted within 1 pdf file of no more than 10 pages.

Applications should be submitted to the following email address: msi@institutfrancaisindia.in mentioning the reference number of the Job offer clearly.

Research Project Title: "Deciphering RNA Evolutionary Landscape by Integrating Directed Evolution and Machine Learning (RE-LEARN)"

Candidates are requested to contact the French scientific principal investigator of the project before submission.
A recommendation letter from the scientific principal investigator is mandatory.

Benefits:

- Monthly allowance of 2400 euros for Post-Doc
- Travel allowance
- University fee
- Carte de séjour fee
- Campus France management fee
- Registration to the French social security scheme

Selection process:

Selection is made by a dedicated selection committee of at least 4 persons. Decisions will be transmitted by the Embassy of France to CEFIPRA. **No consideration will be given for candidates with no recommendation letter from the French institution.**

Criteria for applicants' selection:

Academic excellence

- Excellence of the Academic background, Academic records, Honors, Letters of support, Participation to international research projects, exchange programmes and conferences.

Motivation and qualities

- Academic maturity: appropriation of the thesis project (stakes and contexts) • Quality of the presentation (oral expression, skills for synthesis, English level) • Maturity of the professional project: capacity to project her/himself within five years in terms of career development.

About CEFIPRA:
Indo-French Center for the Promotion of Advanced Research (CEFIPRA/IFCPAR) is an Indian body which promotes scientific cooperation between France and India in advanced fields of Science and Technology. It is supported by the Department of Science and Technology, Government of India and the Ministry of Europe and Foreign Affairs of the French government